Published in final edited form as:
Biometrics. 2011 June ; 67(2): 495–503. doi:10.1111/j.1541-0420.2010.01463.x.

Fixed and Random Effects Selection in Mixed Effects Models

Joseph G. Ibrahim[*], **Hongtu Zhu**[**], **Ramon I. Garcia**[***], and **Ruixin Guo**[****]
Department of Biostatistics, University of North Carolina at Chapel Hill, USA

SUMMARY

We consider selecting both fixed and random effects in a general class of mixed effects models using maximum penalized likelihood (MPL) estimation along with the smoothly clipped absolute deviation (SCAD) and adaptive LASSO (ALASSO) penalty functions. The maximum penalized likelihood estimates are shown to posses consistency and sparsity properties and asymptotic normality. A model selection criterion, called the IC_Q statistic, is proposed for selecting the penalty parameters (Ibrahim, Zhu and Tang, 2008). The variable selection procedure based on IC_Q is shown to consistently select important fixed and random effects. The methodology is very general and can be applied to numerous situations involving random effects, including generalized linear mixed models. Simulation studies and a real data set from an Yale infant growth study are used to illustrate the proposed methodology.

Keywords

ALASSO; Cholesky decomposition; EM algorithm; IC_Q criterion; Mixed Effects selection; Penalized likelihood; SCAD

1. Introduction

In the analysis of mixed effects models, a primary objective is to assess significant fixed effects and/or random effects of the outcome variable. For instance, when simultaneously selecting both random and fixed effects, that is, when selecting *mixed effects*, it is common to use a selection procedure (e.g., forward or backward elimination), coupled with a selection criterion, such as AIC and/or BIC based on the observed data log-likelihood, to compare a set of candidate models (Keselman et al., 1998; Gurka, 2006; Liang, Wu, and Zou, 2008; Ibrahim, Zhu, and Tang, 2008; Claeskens and Consentino, 2008). Zhu and Zhang (2006) proposed a testing procedure based on a class of test statistics for a general mixed effects model to test the homogeneity hypothesis that all of the variance components are zero. Such methods, however, suffer from a serious deficiency in that it is infeasible to simultaneously select significant random and fixed (mixed) effects from a large number of possible models (Fan and Li, 2001; Fan and Li, 2002). To overcome such a deficiency, variable selection procedures based on penalized likelihood methods, such as the Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001) and the Adaptive Lasso (ALASSO) (Zou, 2006), may be developed to select mixed effects.

[*]ibrahim@bios.unc.edu
[**]hzhu@bios.unc.edu
[***]rigarcia@bios.unc.edu
[****]rguo@bios.unc.edu

Supplementary Materials
Web-based supplementary document referenced in Section 3 is available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Compared to the large body of literature on variable selection procedures, we make several novel contributions in this paper. This is one of the few papers on developing selection methods for selecting mixed effects in a large class of mixed effects models. Most variable selection procedures are developed for various parametric models and semiparametric models with/without random effects and/or unobserved data (Fan and Li, 2002, 2004; Cai et al., 2005; Qu and Li, 2006; Zhang and Lu, 2007; Ni, Zhang, and Zhang, 2009; Johnson, Lin, and Zeng, 2008; Garcia, Ibrahim, and Zhu 2010a, 2010b), but all these procedures have only been used for the selection of significant fixed effects. The only exception is the recent work by Krishna (2009) and Bondell, Krishna, and Ghosh (2010), in which only the linear mixed model is considered. We use a novel reparametrization to reformulate the selection of mixed effects into the problem of grouped variable selection in models with heavy 'missing' data, where the missing data is represented by the random effects. This reparametrization makes it possible to use penalized likelihood methods to select both fixed and random effects. Compared to most variable selection methods for linear models, we must address additional challenges due to the presence of missing observations for each subject. A computational challenge here is to directly maximize the observed data log-likelihood function along with the SCAD or ALASSO penalties to select both fixed and random effects and to calculate their estimates. The observed data log-likelihood for complicated mixed effects models is often not available in closed form, and is computationally intractable because it may involve high dimensional integrals which are difficult to approximate. When selecting random effects, this maximization is further complicated because one must eliminate the corresponding row and column of an insignificant random effect and constrain the remaining matrix to be positive definite. Another challenge is to select appropriate penalty parameters in order to produce estimates having proper asymptotic properties (Fan and Li, 2001), whereas existing selection criteria (Kowalchuck et al., 2004; Gurka, 2006; Liang, Wu, and Zou, 2008; Claeskens and Consentino, 2008) are computationally difficult for general mixed effect models.

The goal of this paper is to develop a simultaneous fixed and random effects selection procedure based on the SCAD and ALASSO penalties for application to longitudinal models, correlated models, and/or mixed effects models. We reformulate the problem of selecting mixed effects and develop a method based on the IC_Q criterion to select the penalty parameters. We also specify the penalty parameters in the SCAD and ALASSO penalty functions as a hyperparameter, and then we use the Expectation Maximization (EM) algorithm to simultaneously optimize the penalized likelihood function and estimate the penalty parameters. Under some regularity conditions, we establish the asymptotic properties of the maximum penalized likelihood estimator and the consistency of the IC_Q-based penalty selection procedure.

To motivate the proposed methodolgy, we consider a dataset from a Yale infant growth study (Wasserman and Leventhal, 1993; Stier et al., 1993). The objective of this study is to investigate the relationship between maternal cocaine dependency and child maltreatment (physical abuse, sexual abuse, or neglect). This study had a total of 298 children from the cocaine exposed and unexposed groups. The outcome variable is infant weight (in pounds), which is obtained over several time points. Seven covariates were considered: day of visit, age of mother, gestational age of infant, race, previous pregnancies, gender of infant, and cocaine exposure. Each child had different number and pattern of visits during the study. We consider the mixed effects model by using the seven covariates as fixed effects and the first three covariates as random effects. Our objective in this analysis is to select the most important predictors of infant weight as well as select significant random effects. The selection of random effects is crucial in this application, as it is not at all clear whether a random intercept model will suffice or whether the longitudinal model should also contain random slope effects. Moreover, there is large number of covariates to select from in the

fixed effects component of the model. The selection can be done by our penalized likelihood method, which includes a penalty function (SCAD or ALASSO) with a random effect and an IC_Q penalty estimate. More details regarding the analyses of these data set is given in Section 5.

The rest of the paper is organized as follows. Section 2 introduces the general development for maximizing the penalized likelihood function and selecting the penalty parameters. Section 3 examines the asymptotic properties of the maximum penalized likelihood (MPL) estimator and the IC_Q penalty selection procedure. Section 4 presents a simulation study to examine the finite sample performance of the maximum penalized likelihood estimate. An real data analysis of the Yale infant growth study is given in Section 5. Section 6 concludes the paper with some discussion.

2. Mixed effects selection for mixed effects models

2.1 Model Formulation

Suppose we observe n independent observations $(\mathbf{y}_1, \mathbf{X}_1),\ldots, (\mathbf{y}_n, \mathbf{X}_n)$, where $\mathbf{y}_i$ is an $n_i \times 1$ vector of responses or repeated measures and $\mathbf{X}_i$ is an $n_i \times p$ matrix of fixed covariates for $i = 1,\ldots, n$. We assume independence among the different $(\mathbf{y}_i, \mathbf{X}_i)$'s and

$$E[\mathbf{y}_i|\mathbf{b}_i, \mathbf{X}_i;\theta]=g(\mathbf{X}_i\beta+\mathbf{Z}_i\mathbf{\Gamma}\mathbf{b}_i), \tag{1}$$

where $\mathbf{b}_i$ is a $q \times 1$ vector of unobserved random effects, $\boldsymbol{\theta}$ denotes all the unknown parameters, $\mathbf{\Gamma}$ is a $q \times q$ lower triangular matrix, $g(\cdot)$ is an known link function, $\boldsymbol{\beta} = (\beta_1,\ldots, \beta_p)^T$ is a $p \times 1$ vector of regression coefficients, and $\mathbf{Z}_i$ is an $n_i \times q$ matrix composed of the columns of $\mathbf{X}_i$. In practice, it is common to assume that the conditional distribution of $\mathbf{y}_i$ given $(\mathbf{b}_i, \mathbf{X}_i)$, denoted by $f(\mathbf{y}_i|\mathbf{b}_i, \mathbf{X}_i; \boldsymbol{\theta})$, belongs to the exponential family, such as the binomial, normal, and Poisson (Little and Schluchter (1985), and Ibrahim and Lipsitz (1996)). For notational simplicity, the random effects $\mathbf{b}_i \sim N_q(0, \mathbf{I}_q)$ are assumed to follow a multivariate normal distribution with zero mean and a $q \times q$ covariance matrix $\mathbf{I}_q$. Equivalently, $\mathbf{\Gamma}\mathbf{b}_i \sim N_q(0, \mathbf{D} = \mathbf{\Gamma}\mathbf{\Gamma}^T)$ and $\mathbf{\Gamma}$ is the Cholesky composition of the $q \times q$ matrix $\mathbf{D}$. We allow the possibility of $\mathbf{D}$ being positive semi-definite so that certain components of $\mathbf{\Gamma}\mathbf{b}_i$ may not be random but 0 with probability 1.

2.2 EM Algorithm for Maximizing the Penalized Likelihood

Selecting mixed effects involves identifying the nonzero components of $\boldsymbol{\beta}$, determining the nonrandom elements of $\mathbf{\Gamma}\mathbf{b}_i$, and simultaneously estimating all nonzero parameters. We propose to maximize the penalized likelihood function given by

$$\mathrm{PL}(\theta)=\ell(\theta) - n\sum_{j=1}^{p}\phi_{\lambda_j}(|\beta_j|) - n\sum_{k=1}^{q}\phi_{\lambda_{p+k}}(\|\gamma_k\|), \tag{2}$$

where $\ell(\theta)=\sum_{i=1}^{n}\ell_i(\theta)$, in which $\ell_i(\boldsymbol{\theta}) = \log \int f(\mathbf{y}_i, \mathbf{b}_i|\mathbf{X}_i; \boldsymbol{\theta})\, d\mathbf{b}_i$ is the observed-data log-likelihood for the ith individual, λ_j is the penalty parameter of $\boldsymbol{\beta}_j$, and the penalty function $\phi_{\lambda_j}(\cdot)$ is a nonnegative, nondecreasing, and differentiable function on $(0, \infty)$ (Fan and Li, 2001; Zou, 2006). In addition, the $k \times 1$ vector γ_k consists of all nonzero elements of the k-th row of the lower triangular $q \times q$ matrix $\mathbf{\Gamma}$, $\|\gamma_k\|=(\gamma_k^T\gamma_k)^{1/2}$, and λ_{p+k} is the group penalty parameter corresponding to the whole k-th row of $\mathbf{\Gamma}$. The structure in (2) ensures that certain estimates of $\boldsymbol{\beta}$ are zero (Fan and Li, 2001), which are insignificant predictors of the outcome

variable, and the other covariates are significant predictors. The penalization of γ_k is performed in a group manner in order to preserve the positive definite constraint on $\mathbf{D}$ such that the estimates of the parametric vector γ_k either are all not zero or all equal to zero (Yuan and Li, 2006). If all the elements of γ_k are zero, then the k-th row of $\mathbf{\Gamma}$ is zero and the k-th element of $\mathbf{\Gamma b}_i$ is not random.

Similar to Chen and Dunson (2003), we reparametrize the linear predictor as

$$\mathbf{X}_i\beta+\mathbf{Z}_i\mathbf{\Gamma b}_i=(\mathbf{X}_i(\mathbf{b}_i^T\otimes\mathbf{Z}_i)\mathbf{J}_q)\begin{pmatrix}\beta\\ \gamma\end{pmatrix}=\mathbf{U}_i\delta, \tag{3}$$

where $\mathbf{J}_q$ is the $q^2\times q(q+1)/2$ matrix which transforms γ to vec($\mathbf{\Gamma}$), i.e. vec($\mathbf{\Gamma}$) = $\mathbf{J}_q\gamma$. By reparametrizing the linear predictor this way, the selection of mixed effects is equivalent to the problem of grouped variable selection in regression models with missing covariates, while the random effects in the design matrix $\mathbf{U}_i$ can be interpreted as the "missing covariates". Using this reparametrization, we can apply the variable selection methods proposed in Garcia, Ibrahim, and Zhu (2010a; 2010b) to select important mixed effects.

Because the observed-data log-likelihood function usually involves intractable integration, we develop a Monte Carlo EM algorithm to compute the maximum penalized likelihood estimator of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}_\lambda$, for each $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_{p+q})$. Denote the complete and observed data for subject i by $\mathbf{d}_{c,i}=(\mathbf{y}_i,\mathbf{X}_i,\mathbf{b}_i)$ and $\mathbf{d}_{o,i}=(\mathbf{y}_i,\mathbf{X}_i)$, respectively, and the entire complete and observed data by $\mathbf{d}_c$ and $\mathbf{d}_o$, respectively. At the s-th iteration, given $\boldsymbol{\theta}^{(s)}$, the E step is to evaluate the *penalized Q-function*, given by

$$Q_\lambda(\theta|\theta^{(s)})=\sum_{i=1}^{n}E\{\log f(\mathbf{d}_{i,c};\theta)|\mathbf{d}_o;\theta^{(s)}\}-n\sum_{j=1}^{p}\phi_{\lambda_j}(|\beta_j|)-n\sum_{k=1}^{q}\phi_{\lambda_{p+k}}(||\gamma_k||) \tag{4}$$

$$=Q_1(\theta|\theta^{(s)})-n\sum_{j=1}^{p}\phi_{\lambda_j}(|\beta_j|)-n\sum_{k=1}^{q}\phi_{\lambda_{p+k}}(||\gamma_k||)+Q_2(\theta^{(s)}), \tag{5}$$

where $\boldsymbol{\theta}=(\boldsymbol{\delta}^T,\boldsymbol{\xi}^T)^T$, in which $\boldsymbol{\xi}$ represents all other parameters other than $\boldsymbol{\delta}$, $\mathbf{d}_{i,c}=(\mathbf{y}_i,\mathbf{b}_i,\mathbf{X}_i)$, and

$$Q_1(\theta|\theta^{(s)})=\sum_{i=1}^{n}\int\{\log f(\mathbf{y}_i|\mathbf{b}_i,\mathbf{X}_i;\delta,\boldsymbol{\xi})\}f(\mathbf{b}_i|\mathbf{d}_{i,o};\theta^{(s)})\,d\mathbf{b}_i, \tag{6}$$

$$Q_2(\theta^{(s)})=\sum_{i=1}^{n}\int\{\log f(\mathbf{b}_i)\}f(\mathbf{b}_i|\mathbf{d}_{i,o};\theta^{(s)})\,d\mathbf{b}_i. \tag{7}$$

Since the integrals in (6) and (7) are often intractable, we approximate these integrals by taking a Markov chain Monte Carlo (MCMC) sample of size L from the density $f(\mathbf{b}_i|\mathbf{d}_{i,o};\boldsymbol{\theta}^{(s)})$ (See Ibrahim, Chen, and Lipsitz, 1999). Let $\mathbf{b}_i^{(s,l)}$ be the l-th simulated value at the s-th iteration of the algorithm. The integrals in (6) can be approximated as,

$$Q_1(\theta|\theta^{(s)})=\frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{n}\log f(\mathbf{y}_i|\mathbf{b}_i^{(s,l)},\mathbf{X}_i;\theta). \tag{8}$$

The M step involves maximizing

$$Q_{1,\lambda}(\theta|\theta^{(s)})=Q_1(\theta|\theta^{(s)})-n\sum_{j=1}^{p}\phi_{\lambda_j}(|\beta_j|)-n\sum_{k=1}^{q}\phi_{\lambda_{p+k}}(||\gamma_k||) \tag{9}$$

with respect to $(\boldsymbol{\delta}, \boldsymbol{\xi})$. Maximizing $Q_{1,\lambda}(\boldsymbol{\delta}, \boldsymbol{\xi}|\boldsymbol{\theta}^{(s)})$ with respect to $\boldsymbol{\xi}$ is straightforward and can be done using a standard optimization algorithm, such as the Newton-Raphson algorithm (Little and Schluchter, 1985; Ibrahim, 1990; Ibrahim and Lipsitz, 1996). Maximizing $Q_{1,\lambda}$ with respect to $\boldsymbol{\delta}$ is difficult because $Q_{1,\lambda}$ is a nondifferentiable and nonconcave function of $\boldsymbol{\delta}$ respectively (Zou and Li, 2008).

In order to maximize $Q_{1,\lambda}$, following Fan and Li (2001), a second order Taylor's series approximation of $Q_{1,\lambda}$ centered at the value $\boldsymbol{\delta}^{(s)}$ is used. Using this approximation, $Q_{1,\lambda}$ resembles a penalized weighted least squares regression, so algorithms for minimizing penalized least squares can be used (Fan and Li, 2001; Hunter and Li, 2005). We use a modification of the local linear approximation algorithm (LLA) (Zou and Li, 2008) to incorporate grouped penalization. For γ_k, we use an approximation centered at $\gamma_k^{(s)}$ as follows:

$$\phi_{\lambda_{p+k}}(||\gamma_k||)\approx\sum_{t=1}^{k}\left\{\frac{\phi_{\lambda_{p+k}}(||\gamma_k^{(s)}||)|\gamma_{kt}^{(s)}|}{||\gamma_k^{(s)}||}\right\}|\gamma_{kt}|, \tag{10}$$

where γ_{kt} is the t-th element of the vector γ_k and we assume $||\gamma_k^{(s)}||>0$. If $||\gamma_k^{(s)}||=0$, then we let $\gamma_k^{(s+1)}=0$. Using this approximation, $Q_{1,\lambda}$ resembles a penalized regression with an L_2 penalty, so the methods for performing the lasso can be used to maximize $Q_{1,\lambda}$ (Tibshirani, 1996; Fu, 1998).

Let $\boldsymbol{\xi}^{(s+1)}=\underset{\boldsymbol{\xi}}{\text{argmax}}\ Q_{1,\lambda}(\delta^{(s)},\boldsymbol{\xi}|\theta^{(s)})$ and $\delta^{(s+1)}=\underset{\delta}{\text{argmax}}\ Q_{1,\lambda}(\delta,\boldsymbol{\xi}^{(s+1)}|\theta^{(s)})$. Due to the Taylor's series approximation of Q_1 and the LLA of $\boldsymbol{\phi}_{\lambda_j}$, $\boldsymbol{\theta}^{(s+1)}=(\boldsymbol{\delta}^{(s+1)}, \boldsymbol{\xi}^{(s+1)})$ may not necessarily be the maximizer of $Q_\lambda(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$. By implementing the Expectation Conditional-Maximization (ECM) algorithm (Meng and Rubin, 1993), however, we can find a $\boldsymbol{\theta}^{(s+1)}$ such that $Q_\lambda(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\theta}^{(s)})\geq Q_\lambda(\boldsymbol{\theta}^{(s)}|\theta^{(s)})$ instead of directly maximizing $Q_\lambda(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$. This process is iterated until convergence and the value at convergence, denoted by $\hat{\boldsymbol{\theta}}_\lambda$, maximizes the penalized observed data log likelihood function.

2.3 Penalty Parameter Selection Procedure

To ensure that $\hat{\boldsymbol{\theta}}_\lambda$ has good properties, the penalty parameter $\boldsymbol{\lambda}$ has to be appropriately selected. Two commonly used criteria for selection of the penalty parameter include the Generalized Cross Validation (GCV) and BIC criteria (Wang et al., 2007). These criteria cannot be easily computed in the presence of random effects, because they are functions of observed data quantities whose expressions may require intractable integrals. Moreover, it

has been shown in Wang et al. (2007) that even in the simple linear model, the GCV criterion can lead to significant overfit.

We propose two methods to select the penalty parameter: an IC_Q criterion and a random effects penalty selection method. The IC_Q criterion (Ibrahim, Zhu, and Tang, 2008) selects the optimal $\boldsymbol{\lambda}$ by minimizing

$$IC_Q(\lambda) = -2Q(\widehat{\theta}_\lambda|\widehat{\theta}_0)+c_n(\widehat{\theta}_\lambda),$$

where $\widehat{\theta}_0=\underset{\theta}{\operatorname{argmax}}\,\ell(\theta)$ is the unpenalized maximum likelihood estimate and $c_n(\boldsymbol{\theta})$ is a function of the data and the fitted model. For instance, if c_n equals twice the total number of parameters, then we obtain an AIC-type criterion; alternatively, we obtain a BIC-type criterion when $c_n(\boldsymbol{\theta}) = \dim(\boldsymbol{\theta}) \times \log n$. Moreover, in the absence of random effects, $IC_Q(\boldsymbol{\lambda})$ reduces to the usual AIC or BIC criteria. As in the EM algorithm, we can draw a set of samples from $f(\mathbf{b}_i|\mathbf{d}_{i,o}; \hat{\boldsymbol{\theta}}_0)$ for $i = 1,\ldots, n$ in order to estimate $Q(\hat{\boldsymbol{\theta}}_\lambda|\hat{\boldsymbol{\theta}}_0)$ for any $\boldsymbol{\lambda}$.

The random effects penalty estimator is calculated under the assumption that $\boldsymbol{\delta}$ is distributed as a random effect vector in a hierarchical model. The quantity $\boldsymbol{\lambda}$ can be regarded as a hyperparameter vector in the distribution of $\boldsymbol{\delta}$, denoted by $f(\boldsymbol{\delta}|\boldsymbol{\lambda}, n)$. Then, $\boldsymbol{\lambda}$ can be estimated by maximizing the marginal likelihood with respect to $(\boldsymbol{\xi}, \boldsymbol{\lambda})$, which is given by

$$\int\prod_{i=1}^{n}\int f(y_i|\mathbf{X}_i, \mathbf{b}_i, \delta;\boldsymbol{\xi})f(\mathbf{b}_i)f(\delta|\lambda, n)\,d\mathbf{b}_i\,d\delta=\prod_{i=1}^{n}\int f(\mathbf{d}_{i,o}|\boldsymbol{\xi})f(\delta|\lambda, n)\,d\delta, \tag{11}$$

where $f(\boldsymbol{\delta}|\boldsymbol{\lambda}, n)$ is defined by

$$f(\delta|\lambda, n)=\prod_{j=1}^{p}\exp\{-n\phi_{\lambda_j}(|\beta_j|)\}\prod_{k=1}^{q}\exp\{-n\phi_{\lambda_{p+k}}(\|\gamma_k\|)\}/\{C(\lambda, n)\},$$

and $C(\boldsymbol{\lambda}, n)$ is the normalizing constant of $f(\boldsymbol{\delta}|\boldsymbol{\lambda}, n)$. The resulting estimate of $\boldsymbol{\lambda}$, denoted by $\hat{\boldsymbol{\lambda}}_{RE}$, from the maximization of (11), is the random effects penalty estimator. Treating $\boldsymbol{\delta}$ as missing data, the Monte Carlo EM algorithm can be used to maximize (11) with respect to $(\boldsymbol{\xi}, \boldsymbol{\lambda})$.

We consider the SCAD and ALASSO penalty functions for determining $\boldsymbol{\lambda}$. The ALASSO penalty is defined by

$$\phi_{\lambda_j}(|\beta_j|)=\lambda_j|\beta_j| \text{ for } j=1,\cdots,p, \phi_{\lambda_{p+k}}(\|\gamma_k\|)=\lambda_{p+k}\|\gamma_k\| \text{ for } k=1,\cdots,q.$$

Typical values of λ_j are $\lambda_j = \lambda_{01}|\hat{\beta}_j|^{-1}$ and $\lambda_{p+k}=\lambda_{02}\sqrt{k}\|\widehat{\gamma}_k\|^{-1}$, where $\hat{\beta}_j$ and $\hat{\gamma}_k$ are the unpenalized maximum likelihood (ML) estimates. The multiplier $\sqrt{k}$ normalizes the penalty parameter γ_k in order to accommodate the varying sizes of γ_k. When $\lambda_j = \lambda_{01}$ and $\lambda_{p+k}=\lambda_{02}\sqrt{k}$, the ALASSO reduces to the LASSO penalty.

The SCAD penalty (Fan and Li, 2001) is a nonconcave function defined by $\boldsymbol{\phi}_\lambda(0) = 0$ and for $|\beta|>0, \phi'_\lambda(|\beta|)=\lambda 1(|\beta| \le \lambda)+\frac{(a\lambda-|\beta|)+}{a-1}1(|\beta|>\lambda)$, where t_+ denotes the positive part of t and

a = 3.7. Because the integral of the negative exponential of the ALASSO and SCAD penalties is not finite, i.e. $\int_{-\infty}^{\infty}\exp\{-n\phi_\lambda(\|\lambda_k\|)\}d\gamma_k=\infty$, the expression $\exp\{-n\phi_\lambda(\|\lambda_k\|)\}$ is defined in a bounded space to ensure that $f(\boldsymbol{\delta}|\boldsymbol{\lambda}, n)$ is a proper density. Since a closed form expression of $\hat{\boldsymbol{\lambda}}_{RE}$ is unavailable for both the ALASSO and SCAD penalties, we use the Newton Raphson algorithm along with the ECM algorithm to estimate $\hat{\boldsymbol{\lambda}}_{RE}$.

3. Theoretical Results

In this section, we establish the asymptotic theory of the MPL estimator and the consistency of the penalty selection procedure based on IC_Q. Suppose $\beta=\left(\beta_{(1)}^T,\beta_{(2)}^T\right)^T$, where $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$ are, respectively, $p_1 \times 1$ and $(p - p_1) \times 1$ subvectors. Let $\beta^*=\left(\beta_{(1)}^{*T},\beta_{(2)}^{*T}\right)^T$ denote the true value of $\boldsymbol{\beta}$. Without loss of generality, we assume that $\beta_{(2)}^*=0$ and all of the components of $\beta_{(1)}^*$ are not equal to zero. Similarly let $\gamma=\left(\gamma_1^T,\ldots,\gamma_k^T\right)^T=\left(\gamma_{(1)}^T,\gamma_{(2)}^T\right)^T$ where $\gamma_{(1)}^T=\left(\gamma_1^T,\ldots,\gamma_{q_1}^T\right)^T$, $\gamma_{(2)}^T=\left(\gamma_{q_1+(1)}^T,\ldots,\gamma_q^T\right)^T$ and $\gamma_{(1)}$ and $\gamma_{(2)}$ are $q_1(q_1 + 1)/2 \times 1$ and $\{q - q_1(q_1 + 1)/2\} \times 1$ subvectors respectively. Let $\gamma^*=\left(\gamma_{(1)}^{*T},\gamma_{(2)}^{*T}\right)^T$ denote the true value of γ. Without loss of generality, we assume that $\gamma_{(2)}^*=0$ and some of the components of each γ_k^* are not equal to zero for $k = 1,\ldots, q_1$.

Let $\mathcal{S}=\{j_{11},\ldots,j_{1d_1};j_{21},\ldots,j_{2d_2}\}$ be a candidate model containing the j_{11}-th, …, j_{1d}-th columns of $\mathbf{X}$ and the j_{21}-th, …, j_{2d_2}-th columns of $\mathbf{Z}$. Thus, $\mathcal{S}_F=\{1,\ldots,p;1,\ldots,q\}$ and $\mathcal{S}_T=\{1,\ldots,p_1;1,\ldots,q_1\}$ denote the full and true covariate models, respectively. If $\mathcal{S}$ misses at least one important covariate, that is $\mathcal{S}\not\supset\mathcal{S}_T$, then $\mathcal{S}$ is referred to as an underfitted model; however, if $\mathcal{S}\supset\mathcal{S}_T$ and $\mathcal{S}\neq\mathcal{S}_T$, then $\mathcal{S}$ is an overfitted model. The unpenalized and penalized ML estimators of $\boldsymbol{\theta}=(\boldsymbol{\beta}^T,\gamma^T,\xi)^T$, denoted by $\hat{\theta}_S$ and $\hat{\boldsymbol{\theta}}_\lambda$, respectively, are defined as

$$\widehat{\theta}_S=\underset{\theta:\beta_j\neq 0,\forall\, j\in\mathcal{S}}{\operatorname{argmax}}\ \ell(\theta)\ \text{ and }\widehat{\theta}_\lambda=\underset{\theta}{\operatorname{argmax}}\left\{\ell(\theta)-n\sum_{j=1}^{p}\phi_{\lambda_j}(|\beta_j|)-n\sum_{k=1}^{q}\phi_{\lambda_{p+k}}(\|\gamma_k\|)\right\},$$

, and particularly $\hat{\boldsymbol{\theta}}_{\mathcal{S}_F}=\hat{\boldsymbol{\theta}}_0$. We obtain the following theorems whose assumptions and proofs can be found in the web-based supplementary document.

THEOREM 1

Under assumptions (C1)–(C7) in the supplementary document, *we have*

- **a.** $\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}^*=O_p(n^{-1/2})$ *as* $n\to\infty$, *where* $\boldsymbol{\theta}^*$ *is the true value of* $\boldsymbol{\theta}$;
- **b.** *Sparsity:* $P(\hat{\boldsymbol{\beta}}_{(2)\lambda}=0,\hat{\gamma}_{(2)\lambda}=0)\to 1$;
- **c.** *Asymptotic normality:* $\sqrt{n}\{(\widehat{\beta}_{(1)\lambda}^T,\widehat{\lambda}_{(1)\lambda}^T,\widehat{\boldsymbol{\xi}}_\lambda^T)^T-(\beta_{(1)}^{*T},\gamma_{(1)}^{*T},\boldsymbol{\xi}_{(1)}^{*T})^T\}$ *is asymptotically normal with mean and covariance matrix defined in the supplement.*

Theorem 1 states that by appropriately choosing the penalty $\boldsymbol{\lambda}$, there exists a root-n estimator of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_\lambda$, and that this estimator must possess the sparsity property, i.e. $\hat{\boldsymbol{\beta}}_{(2)\lambda}=0$, $\hat{\gamma}_{(2)\lambda}=0$ in probability. Moreover, $(\widehat{\beta}_{(1)\lambda}^T,\widehat{\lambda}_{(1)\lambda}^T,\widehat{\boldsymbol{\xi}}_\lambda^T)^T$ is asymptotically normal.

We investigate whether the $IC_Q(\boldsymbol{\lambda})$ criterion can consistently select the correct model. For each $\boldsymbol{\lambda} \in R^{p+}$, $(\hat{\boldsymbol{\beta}}_\lambda, \hat{\gamma}_\lambda)$ naturally defines a candidate model $\mathbb{S}_\lambda = \{j : \hat{\boldsymbol{\beta}}_{\lambda,j} \neq 0; k : \|\hat{\gamma}_{\lambda,k}\| \neq 0\}$. Generally, $\mathcal{S}_\lambda$ can be either underfitted, overfitted, or true. Therefore, R^{p+} can be partitioned into three mutually exclusive regions

$R_u^{p+}=\{\lambda \in R^{p+}:\mathcal{S}_\lambda \not\supset \mathcal{S}_T\}, R_t^{p+}=\{\lambda \in R^{p+}:\mathcal{S}_\lambda=\mathcal{S}_T\},$ and $R_o^{p+}=\{\lambda \in R^{p+}:\mathcal{S}_\lambda \not\supset \mathcal{S}_T, \mathcal{S}_\lambda \neq \mathcal{S}_T\},$. Furthermore, if we can choose a reference penalty parameter sequence $\{\lambda_n \in R^{p+}\}_{n=1}^{\infty}$, which satisfies the conditions of Theorem 1, then $\mathcal{S}_{\boldsymbol{\lambda}_n} = \mathcal{S}_T$ in probability.

To select $\boldsymbol{\lambda}$ we first calculate

$$\mathrm{dIC}_Q(\lambda_2,\lambda_1)=\mathrm{IC}_Q(\lambda_2) - \mathrm{IC}_Q(\lambda_1)= - 2Q(\widehat{\theta}_{\lambda_2}|\widehat{\theta}_0)+c_n(\widehat{\theta}_{\lambda_2})+2Q(\widehat{\theta}_{\lambda_1}|\widehat{\theta}_0) - c_n(\widehat{\theta}_{\lambda_1})$$

for any two $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$. We assume $\mathcal{S}_{\lambda 2} \supset \mathcal{S}_{\lambda 1}$ and choose the model $\mathcal{S}_{\lambda 1}$ resulting from using the penalty value $\boldsymbol{\lambda}_1$ if $\mathrm{dIC}_Q(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) \geq 0$, otherwise we choose the model $\mathcal{S}\lambda_2$.

Define $\delta_Q(\lambda_1, \lambda_2)=E\{Q(\theta^*_{\mathcal{S}_{\lambda_1}}|\theta^*)\} - E\{Q(\theta^*_{\mathcal{S}_{\lambda_2}}|\theta^*)\}$, and $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) = c_n(\hat{\boldsymbol{\theta}}_{\lambda 2}) - c_n(\hat{\boldsymbol{\theta}}_{\lambda 1})$, where $\theta^*_{\mathcal{S}}$ is defined in the supplementary document.

THEOREM 2

Under assumptions (C1)–(C7) in the supplementary document, *we have the following results.*

- **a.** *If for all* $\mathcal{S}_\lambda \not\supset \mathcal{S}_T$, $\liminf_n \delta_Q(\lambda, 0)/n>0$ *and* $\delta_c(\boldsymbol{\lambda}, 0) = o_p(n)$, *then* $dIC_Q(\boldsymbol{\lambda}, 0) > 0$ *in probability.*
- **b.** *If* $E\{Q(\theta^*_{\mathcal{S}_{\lambda_1}}|\widehat{\theta}_0)\}-E\{Q(\theta^*_{\mathcal{S}_{\lambda_2}}|\widehat{\theta}_0)\}=O_p(n^{1/2})$ *and* $Q(\widehat{\theta}_{\lambda_t}|\widehat{\theta}_0)-E\{Q(\theta^*_{\mathcal{S}_{\lambda_t}}|\widehat{\theta}_0)\}=O_p(n^{1/2})$ *for* $t = 1, 2$, *then* $dIC_Q(\boldsymbol{\lambda}_2, \lambda_1) > 0$ *in probability as* $n^{-1/2}\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1)$ *converges to* ∞ *in probability.*
- **c.** *If* $Q(\hat{\boldsymbol{\theta}}_{\lambda_1}|\hat{\boldsymbol{\theta}}_0) - Q(\hat{\boldsymbol{\theta}}_{\lambda_2}|\hat{\boldsymbol{\theta}}_0) = O_p(1)$, *then* $dIC_Q(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) > 0$ *in probability as* $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1)$ *converges to* ∞ *in probability.*

Theorem 2 has some important implications. Theorem 2(a) shows that $IC_Q(\boldsymbol{\lambda})$ chooses all significant covariates with probability 1. Because $\mathcal{S}_0 \subset R_t^p \cup R_o^p$, the optimal model selected by minimizing $IC_Q(\boldsymbol{\lambda})$ will not select a $\boldsymbol{\lambda}$ with $\mathcal{S}_\lambda \not\supset \mathcal{S}_T$ because $\mathrm{dIC}_Q(\boldsymbol{\lambda}, 0) > 0$ in probability. Therefore, the $IC_Q(\boldsymbol{\lambda})$ criterion selects all significant covariates with probability tending to 1. Generally, the most commonly used $c_n(\boldsymbol{\theta})$, such as $2\dim(\boldsymbol{\theta})$, $\dim(\boldsymbol{\theta})\log(n)$, and $K \log\log(n)$ $(K > 0)$, satisfy the condition $\delta_c(\boldsymbol{\lambda}, 0) = o_p(n)$. The condition $\liminf_n n^{-1}\delta_Q(\lambda, 0)>0$ ensures that $IC_Q(\boldsymbol{\lambda})$ chooses a model with large $E\{Q(\theta^*_{\mathcal{S}}|\theta^*)\}$. This condition is analogous to condition 2 in (Wang et al., 2007), which elucidates the effect of underfitted models. The term $n^{-1}E\{Q(\theta^*|\theta^*)\} - n^{-1}E\{Q(\theta^*_{\mathcal{S}}|\theta^*)\}$ can be written as

$$n^{-1}\ell(\theta^*) - n^{-1}\ell(\theta^*_s)+n^{-1}E\{H(\theta^*|\theta^*)\} - n^{-1}E\{H(\theta^*_s|\theta^*)\}, \tag{12}$$

where

$$H(\theta_1|\theta_2)=\sum_{i=1}^{n}\int \log\{f(\mathbf{b}_i|\mathbf{d}_{o,i};\theta_1)\}f(\mathbf{b}_{im}|\mathbf{d}_{o,i};\theta_2)\, d\mathbf{b}_{im}. \tag{13}$$

By Jensen's inequality, the third and fourth terms of (12) are greater than zero and the first and second terms must be greater than zero for large n. Thus, $\lim\inf_n n^{-1}\delta_Q(\boldsymbol{\lambda}, 0) \geq 0$ in probability.

If $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ have the same average $n^{-1}E\{Q(\theta^*_{\mathscr{S}_\lambda}|\theta^*)\}$, that is, $\lim\inf_n n^{-1}\delta_Q(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) = 0$, then Theorem 2 (b) and (c) indicate that $\mathrm{IC}_Q(\boldsymbol{\lambda})$ picks out the smaller model $\mathscr{S}_{\lambda_1}$ when $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1)$ increases to ∞ at a certain rate (e.g., $\log(n)$). For example, for the BIC-type criterion, $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) = \{\dim(\hat{\boldsymbol{\theta}}_{\mathscr{S}_{\lambda_2}}) - \dim(\hat{\boldsymbol{\theta}}_{\mathscr{S}_{\lambda_1}})\}\log(n) \geq \log(n)$ since we assume $\mathscr{S}_{\lambda_2} \supset \mathscr{S}_{\lambda_1}$. The AIC-type criterion, for which $c_n(\boldsymbol{\theta}) = 2 \times \dim(\boldsymbol{\theta})$, however, does not satisfy this condition. Thus, similar to the AIC criterion with no random effects, $\mathrm{IC}_Q(\boldsymbol{\lambda})$ with $c_n(\boldsymbol{\theta}) = 2 \times \dim(\boldsymbol{\theta})$ tends to overfit.

4. Simulation Study

We use simulations to examine the finite sample performance of the maximum penalized likelihood estimates using our proposed penalty estimators and compare them to the unpenalized ML estimate. Our objectives for these simulations are to 1) compare the random effects and IC_Q penalty estimators and 2) to compare the SCAD, LASSO, and ALASSO penalty functions.

To do this, we simulated a data set consisting of n independent observations according to the model $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\Gamma}\mathbf{b}_i + \sigma\boldsymbol{\varepsilon}_i$, $i = 1,\ldots, n$, where $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_i$ are independent and standard multivariate normal random vectors, and $\boldsymbol{\beta} = (3, 2, 1.5, 0, 0, 0, 0, 0)^T$. Moreover, $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \mathbf{D}$ is a 3×3 matrix, such that the (r, s) element of $\mathbf{D}$ is $\rho^{|r-s|}$. The matrix $\mathbf{X}_i$ is a 12×8 matrix of independent rows, where each row of $\mathbf{X}_i$ has mean zero and covariance matrix Σ_{xx} whose (r, s) element is $\rho^{|r-s|}$. The matrix $\mathbf{Z}_i$ was set equal to $\mathbf{X}_i$.

We considered six different settings: ($n = 50$, $\sigma = 3$), ($n = 50$, $\sigma = 1$), ($n = 100$, $\sigma = 3$), ($n = 100$, $\sigma = 1$), ($n = 200$, $\sigma = 3$), and ($n = 200$, $\sigma = 1$) with a value of $\rho = .5$ for all settings. For each setting, one design matrix was simulated and 100 data sets $(\mathbf{y}_i, \mathbf{X}_i)$ for $i = 1,\ldots, n$ were generated.

For each simulated data set, the maximum penalized likelihood (MPL) estimate using the SCAD, LASSO and ALASSO penalties was computed using the random effects and IC_Q penalty estimates. These estimates are denoted as SCAD-RE, SCAD-IC_Q, LASSO-RE, LASSO-IC_Q, ALASSO-RE, and ALASSO-IC_Q, respectively. For the IC_Q estimate, the BIC-type criterion, $c_n(\boldsymbol{\theta}) = \dim(\boldsymbol{\theta})\log n$, was used. For the Monte Carlo EM algorithm, 2000 Monte Carlo iterations were used within each iteration of EM. For the SCAD and LASSO penalties, we set $\lambda_j = \lambda_{01}$, for $j = 1,\ldots 8$, and $\lambda_{8+k} = \lambda_{02}\sqrt{k}$, for $k = 1,\ldots,3$ while for the ALASSO penalty, $\lambda_j = \lambda_{01}|\hat{\beta}_j|^{-1}$, for $j = 1,\ldots 8$, and $\lambda_{8+k} = \lambda_{02}\sqrt{k}\|\hat{\gamma}_k\|^{-1}$ for $k = 1,\ldots,3$ where $\hat{\beta}_j$, and $\hat{\gamma}_k$ are the unpenalized ML estimates of β_j and γ_k respectively, and the penalty $(\lambda_{01}, \lambda_{02})$ was estimated using the IC_Q and random effects penalty selection methods.

For each estimate, the penalized estimate of $\boldsymbol{\beta}$ and $\mathbf{D}$ were computed, denoted as $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\mathbf{D}}_\lambda$ respectively, and the mean square error $\mathrm{ME}(\hat{\boldsymbol{\beta}}_\lambda) = (\hat{\boldsymbol{\beta}}_\lambda - \beta)^T\Sigma_{xx}(\hat{\boldsymbol{\beta}}_\lambda - \beta)^T$ and the quadratic loss error $\mathrm{ME}(\hat{\mathbf{D}}_\lambda) = \mathrm{trace}[(\hat{\mathbf{D}}_\lambda - \mathbf{D})^2]^{1/2}$ were computed. The ratio of the model error of the MPL estimate to that of the unpenalized ML estimate, $\mathrm{ME}(\hat{\boldsymbol{\beta}}_\lambda)/\mathrm{ME}(\hat{\boldsymbol{\beta}}_0)$ and $\mathrm{ME}(\hat{\mathbf{D}}_\lambda)/\mathrm{ME}(\hat{\mathbf{D}}_0)$, were computed for each data set and the median of the ratios over the 100 simulated data sets, denoted as MRME, was calculated. The MRME of the true model is also reported. In addition, we report two types of errors regarding the fixed and random effects. ZERO_1 is the mean number of type I errors (an effect is truly not significant or random but the corresponding MPL estimate indicates it is significant or random) and

$ZERO_2$ is the mean number of type II errors (an effect is truly significant or random but the corresponding MPL estimate indicates it is not significant or random).

For the MPL estimates, MRME values greater than one indicate that the estimate performs worse than the ML estimate, values near one indicate it performs as good as the ML estimate, while values near the 'true' MRME value indicate optimal performance. The values $ZERO_1$ and $ZERO_2$ can be interpreted as estimates of the probability of overfit and underfit, respectively, and the value $1 - ZERO_1 - ZERO_2$ is an estimate of the probability of selecting the true model. Ideally, one would like to have MPL estimates with small $ZERO_1$ and $ZERO_2$ values and small MRME values. Overall, the MRME values of all of the MPL estimates were less than or equal to one, which indicates that regardless of the sample size or noise level, the MPL estimates perform better than the ML estimate. Across all samples sizes and noise levels, the MRME values of the MPL estimates using the random effects penalty estimates was higher than the MPL estimates using the IC_Q penalty estimates. For the IC_Q MPL estimates, as the noise level decreases from $\sigma = 3$ to $\sigma = 1$, the MRME values increase. For a fixed noise level, the MRME values at sample sizes of $n = 50$ and $n = 200$ are comparable but there is a slight decrease in the MRME values at sample sizes of $n = 100$. This indicates that the MPL estimates perform better, relative to the MLE, at low noise levels and near sample sizes of $n = 100$. The MPL estimates using the random effects penalty estimate tended to overfit significantly. On average, the MPL estimate using the ALASSO penalty function had smaller estimation error and overfit than the LASSO estimate. For estimating fixed effects, the SCAD-IC_Q estimate has, on average, smaller estimation error and overfit than the other estimates. For estimating the random effects, the ALASSO-IC_Q has smaller error and overfit.

5. Yale Infant Growth Study

We applied the proposed methodology to the Yale infant growth study of Wasserman and Leventhal (1993) and Stier et al. (1993). The Yale infant growth data were collected to study whether cocaine exposure during pregnancy leads to the maltreatment of infants after birth, such as physical and sexual abuse. A total of 298 children were recruited from two subject groups (cocaine exposure group and unexposed group). Throughout the study different children had different numbers and patterns of visits during the study period. The multivariate response was weight of the infant at each visit. Let y_{ij} denote the weight (in pounds) at the j-th visit of infant i, for $i = 1,\ldots, 298$, $j = 1,\ldots, n_i$ and let $\mathbf{y}_i = (y_{i1},\ldots, y_{in_i})$. The covariates used were: x_{ij1} = day of visit, x_{ij2} = age (in years) of mother, x_{ij3} = gestational age (in weeks) of infant, x_{ij4} = race (2 levels: African American and other, coded as 1 and 0), x_{ij5} = previous pregnancies (2 levels: no and yes, coded as 1 and 0), x_{ij6} = gender of infant (2 levels: male and female, coded as 1 and 0), x_{ij7} = cocaine exposure (2 levels: yes and no, coded as 1 and 0). The design matrix $\mathbf{X}_i$ is a $n_i \times 8$ matrix with the j-th row equal to $(1, x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, x_{ij7})$, $\mathbf{Z}_i$ is a $n_i \times 3$ matrix composed of the first 3 continuous covariates of $\mathbf{X}_i$, i.e., the j-th row of $\mathbf{Z}_i$ is $(x_{ij1}, x_{ij2}, x_{ij3})$, and therefore $q = 3$ here. All covariates were centered in the analysis for numerical stability. Further, we assume that $[\mathbf{y}_i|\mathbf{X}_i; \boldsymbol{\beta}, \mathbf{D}]$ is normally distributed with mean $E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\Gamma}\mathbf{b}_i$, where $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \mathbf{D}$ and $[y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}, \mathbf{D}]$ and $[y_{ij'}|\mathbf{X}_i; \boldsymbol{\beta}, \mathbf{D}]$ are independent for $j \neq j'$.

The objective of this analysis was to determine the significant predictors of infant weight and the significant random effects. Because the ALASSO penalty outperformed the LASSO penalty in the simulations, only the SCAD and ALASSO penalty functions were used along with the IC_Q and random effects penalty estimates. Note that the intercept term was not penalized. For the SCAD, $\lambda_j = \lambda_{01}$ for $j = 2,\ldots, 8$ and $\lambda_{8+k} = \lambda_{02}\sqrt{k}$, for $k = 1,\ldots, 3$, while for the ALASSO penalty, $\lambda_j = \lambda_{01}|\hat{\beta}_j|^{-1}$ for $j = 2, \ldots, 8$ and $\lambda_{8+k} = \lambda_{02}\sqrt{k}\|\widehat{\gamma}_k\|^{-1}$, for $k = 1,\ldots, 3$,

where $\hat{\beta}_j$ and $\hat{\gamma}_k$ are the unpenalized ML estimates of β_j and γ_k, respectively, and $(\lambda_{01}, \lambda_{02})$ was estimated using the IC_Q and random effects penalty selection methods.

The results of the analysis are presented in Table 2. The MPL estimates using the SCAD penalty identify visit, gestational age of infant, gender of infant and cocaine exposure as significant predictors of infant weight, and visit as significant random effect. These estimates coincide with the results of the maximum likelihood analysis which identify the same fixed and random effects as significant (significant effects by MLE analysis are indicated by a * in Table 2). The results of using the SCAD with two different sets of penalty estimates are similar. Although the estimates using SCAD with the IC_Q penalty estimates do not shrink the random-effect variances for age and gestational age to 0, these variance estimates are relatively smaller than that of the visit random effect, which still identifies the correct random-effect. The MPL estimate using the ALASSO penalty shrunk two more coefficients of the fixed effects to zero: gender and cocaine. Although these two effects are identified as significant in the MLE, we see that their corresponding MLE estimates are smaller relative to the other significant fixed effects. The estimates using the ALASSO penalty with the IC_Q penalty estimates are close to that of the RE penalty estimates. The MPL estimates using the ALASSO penalty identify visit and gestational age of infant as significant fixed effects, and visit as a significant random effect.

6. Discussion

We have proposed a general method which performs simultaneous fixed and random effects selection as well as estimation. Under certain regularity conditions and appropriate assumptions on the penalty parameters, the maximum penalized likelihood estimate possesses oracle properties. We have used two methods for estimating the penalty parameters, the random effects and IC_Q penalty selection methods, and showed that under an appropriate choice of $c_n(\boldsymbol{\theta})$, the IC_Q penalty estimate chooses all the significant fixed and random effects with probability 1. Since penalized likelihood methods have been shown to perform poorly in finite samples, simulations were performed to examine the finite sample properties of the maximum penalized likelihood estimators and the performance of the Monte Carlo EM algorithm. In the simulations, the SCAD and ALASSO penalty functions using the IC_Q penalty estimate performed best and had significantly less estimation error than the maximum likelihood estimate. Unlike previous implementations of the random effects penalty estimate (Garcia, Ibrahim, and Zhu, 2010a, 2010b), the simulations and real data analysis results show that for mixed effects regression models, the random effects penalty estimate has significant overfit. For estimating fixed effects, the SCAD-IC_Q estimate had, on average, smaller estimation error and overfit, while for estimating random effects, the ALASSO-IC_Q had smaller error and overfit.

Many aspects of this work warrant further research and investigation. Recent developments have shown that there may be more than one plausible scheme for formulating the grouped penalty in the penalized likelihood (Zhao et al., 2009; Breheny and Huang, 2009). To select significant random effects using a cholesky parametrization of the covariance matrix of the random effects requires that each row of the cholesky matrix to be penalized as a group. Other parameters, however, can be grouped and penalized in various ways. For instance, it is possible to group parameters corresponding to the fixed effects if one is interested in determining whether a particular group of fixed effects is significant or not. It is also possible to use different penalty functions for each group of parameters.

The objective of this paper was to perform simultaneous selection of fixed and random effects. To the best of our knowledge, this is the first paper to propose this type of methodology. In the existing literature, (Gurka, 2006; Chen and Dunson, 2003; Daniels and

Kass, 1999, 2001), the predominant approach to mixed effects selection has been to fix either the mean model or the covariance structure of the random effects and then either test variance components or perform variable selection on the mean model (Keselman et al., 1998). This approach, since it fixes certain parts of the model, makes assumptions regarding the model structure which may not inappropriate. A possible reason that simultaneous mixed effects selection may not have been pursued before is perhaps due to the numerical complexity inherent in the model fitting algorithms. With penalized likelihood methods, however, simultaneous mixed effects selection is straightforward to implement and no assumptions are necessary regarding any part of the model.

As it stands, calculating the IC_Q penalty estimator is slightly demanding. An alternative to IC_Q penalty parameter estimation is to select the penalty parameter which optimizes other criteria developed in mixed effects models such as those in Claeskens and Consentino (2008) and Liang, Wu, and Zou (2008). We will formally study these issues in future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank the editor, associate editor and two referees for helpful comments and suggestions, which have led to an improvement of this article. This research was partially supported by NSF grant BCS-08-26844 and NIH grants GM 70335, CA 74015, RR025747-01, MH086633, AG033387, and P01CA142538-01.

References

Bondell HD, Krishna A, Ghosh SK. Joint variable selection for fixed and random effects in linear mixed-effects models. Biometrics. 2010 in press.

Breheny P, Huang J. Penalized methods for bi-level variable selection. Statistics and its Interface. 2009; 2:369–380. [PubMed: 20640242]

Cai J, Fan J, Li R, Zhou H. Variable selection for multivariate failure time data. Biometrika. 2005; 92:303–316. [PubMed: 19458784]

Claeskens G, Consentino F. Variable selection with incomplete covariate data. Biometrics. 2008; 64:1062–1096. [PubMed: 18371121]

Chen Z, Dunson D. Random effects selection in linear mixed models. Biometrics. 2003; 59:762–769. [PubMed: 14969453]

Daniels MJ, Kass RE. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. Journal of the American Statistical Association. 1999; 94:1254–1263.

Daniels MJ, Kass RE. Shrinkage estimators for covariance matrices. Biometrics. 2001; 57:1173–1184. [PubMed: 11764258]

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. Annals of Statistics. 2002; 30:74–99.

Fan J, Li R. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. Journal of American Statistical Association. 2004; 99:710–723.

Fu W. Penalized regressions: The bridge versus the lasso. Journal of Computational and Graphical Statistics. 1998; 7:375–384.

Garcia RI, Ibrahim JG, Zhu H. Variable selection for regression models with missing data. Statistica Sinica. 2010a; 20:149–165. [PubMed: 20336190]

Garcia RI, Ibrahim JG, Zhu H. Variable selection in the Cox regression model with covariates missing at random. Biometrics. 2010b; 66:97–104. [PubMed: 19459831]

Gurka MJ. Selecting the best linear mixed model under REML. American Statistician. 2006; 60:19–26.

Hunter DR, Li R. Variable selection using MM algorithms. Annals of Statistics. 2005; 33:1617–1642. [PubMed: 19458786]

Ibrahim JG. Incomlete data in generalized linear models. Journal of the American Statistical Association. 1990; 85:765–769.

Ibrahim JG, Chen MH, Lipsitz SR. Monte Carlo EM for missing covariates in parametric regression models. Biometrics. 1999; 55:591–596. [PubMed: 11318219]

Ibrahim JG, Lipsitz SR. Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. Biometrics. 1996; 52:1071–1078. [PubMed: 8805768]

Ibrahim JG, Zhu H, Tang N. Model selection criteria for missing-data problems using the em algorithm. Journal of the American Statistical Association. 2008; 103:1648–1658. [PubMed: 19693282]

Johnson B, Lin DY, Zeng D. Penalized estimating functions and variable selection in semiparametric regression models. Journal of the American Statistical Assoication. 2008; 103:672–680.

Keselman HJ, Algina J, Kowalchuk RK, Wolfinger RD. A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. Communications in Statistics - Simulation and Computation. 1998; 27:591–604.

Kowalchuck RK, Keselman HJ, Algina J, Wolfinger RD. The analysis of repeated measurements with mixed-model adjusted F tests. Educational and Psychological Measurement. 2004; 64:224–242.

Krishna, A. North Carolina State University; 2009. Joint variable selection of fixed and random effects in linear mixed-effects model and its oracle properties. unpublished thesis

Leeb H, Potscher BM. Sparse estimators and the oracle property, or the return of Hodges' Estimator. Journal of Econometrics. 2008; 142:201–211.

Liang H, Wu H, Zou G. A note on conditional AIC for linear mixed effects-models. Biometrika. 2008; 95:773–778. [PubMed: 19122890]

Lin X. Variance component testing in generalized linear models with random effects. Biometrika. 1997; 84:309–326.

Little RJA, Schluchter M. Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika. 1985; 72:497–512.

Meng XL, Rubin DB. Maximum likleihood estimation via the ECM algorithm: a general framework. Biometrika. 1993; 80:267–278.

Ni X, Zhang D, Zhang H. Variable selection for semiparametric mixed models in longitudinal studies. Biometrics. 2009; 66:79–88. [PubMed: 19397585]

Qu A, Li R. Quadratic inference functions for varying-coefficient models with longitudinal data. Biometrics. 2006; 62:379–391. [PubMed: 16918902]

Stier DM, Leventhal JM, Berg AT, Johnson L, Mezger J. Are children born to young mothers at increased risk of maltreatment? Pediatrics. 1993; 91:642–648. [PubMed: 8441574]

Thall PF, Vail SX. Some covariance models for longitudinal dount data with overdispersion. Biometrics. 1990; 46:657–671. [PubMed: 2242408]

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B. 1996; 58:267–288.

Wasserman DR, Leventhal JM. Maltreatment of children born to cocaine-dependent mothers. American J. Diseases of Children. 1993; 147:1324–1328.

Wang H, Li R, Tsai CL. Tuning parameter selector for the smoothly clippped absolute deviation method. Biometrika. 2007; 94:553–568. [PubMed: 19343105]

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J. R. Statistic. Soc. B. 2006; 68:49–67.

Zhang H, Lu W. Adaptive-LASSO for Cox's proportional hazards model. Biometrika. 2007; 94:1–13.

Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. Annals of Statistics. 2009; 37:3468–3497.

Zhu HT, Zhang HP. Generalized score test for homogeneity for mixed effects models. Annals of Statistics. 2006; 34:1545–1569.

Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

Zou H, Li R. One-step sparse estimates in noncancave penalized likelihood models. Annals of Statistics. 2008; 36:1509–1533. [PubMed: 19823597]

Table 1

Simulation results of linear mixed effects models comparing SCAD, LASSO, and ALASSO penalty functions with random effect and IC_Q penalty estimates

		β Estimate (D Estimate)		
Model	**Method**	**MRME**	**$ZERO_1$**	**$ZERO_2$**
$n = 50$, $\sigma = 3$	SCAD-RE	0.576 (0.980)	0.11 (0.94)	0.00 (0.00)
	SCAD-IC_Q	0.552 (0.259)	0.01 (0.09)	0.00 (0.01)
	LASSO-RE	0.983 (0.988)	0.99 (1.00)	0.00 (0.00)
	LASSO-IC_Q	0.605 (0.241)	0.04 (0.10)	0.00 (0.01)
	ALASSO-RE	0.949 (0.983)	0.80 (1.00)	0.00 (0.00)
	ALASSO-IC_Q	0.597 (0.263)	0.01 (0.13)	0.00 (0.01)
	True	0.559 (0.228)	0.00 (0.00)	0.00 (0.00)
$n = 50$, $\sigma = 1$	SCAD-RE	0.906 (0.803)	0.58 (1.00)	0.00 (0.00)
	SCAD-IC_Q	0.869 (0.461)	0.03 (0.13)	0.00 (0.00)
	LASSO-RE	0.997 (0.996)	0.99 (1.00)	0.00 (0.00)
	LASSO-IC_Q	0.884 (0.438)	0.04 (0.08)	0.00 (0.00)
	ALASSO-RE	0.983 (0.989)	0.81 (1.00)	0.00 (0.00)
	ALASSO-IC_Q	0.858 (0.441)	0.03 (0.10)	0.00 (0.00)
	True	0.846 (0.439)	0.00 (0.00)	0.00 (0.00)
$n = 100$, $\sigma = 3$	SCAD-RE	0.571 (0.970)	0.13 (0.93)	0.00 (0.00)
	SCAD-IC_Q	0.565 (0.219)	0.01 (0.04)	0.00 (0.00)
	LASSO-RE	0.993 (0.994)	0.99 (1.00)	0.00 (0.00)
	LASSO-IC_Q	0.584 (0.232)	0.01 (0.04)	0.00 (0.00)
	ALASSO-RE	0.949 (0.987)	0.81 (1.00)	0.00 (0.00)
	ALASSO-IC_Q	0.574 (0.205)	0.01 (0.04)	0.00 (0.00)
	True	0.513 (0.196)	0.00 (0.00)	0.00 (0.00)
$n = 100$, $\sigma = 1$	SCAD-RE	0.895 (0.803)	0.57 (1.00)	0.00 (0.00)
	SCAD-IC_Q	0.820 (0.452)	0.01 (0.07)	0.00 (0.00)
	LASSO-RE	0.999 (0.997)	0.99 (1.00)	0.00 (0.00)
	LASSO-IC_Q	0.835 (0.478)	0.03 (0.08)	0.00 (0.00)
	ALASSO-RE	0.982 (0.989)	0.82 (1.00)	0.00 (0.00)
	ALASSO-IC_Q	0.839 (0.415)	0.02 (0.06)	0.00 (0.00)
	True	0.832 (0.392)	0.00 (0.00)	0.00 (0.00)
$n = 200$, $\sigma = 3$	SCAD-RE	0.553 (0.987)	0.13 (0.94)	0.00 (0.00)
	SCAD-IC_Q	0.554 (0.245)	0.01 (0.07)	0.00 (0.00)
	LASSO-RE	0.995 (0.996)	0.99 (1.00)	0.00 (0.00)
	LASSO-IC_Q	0.617 (0.244)	0.05 (0.09)	0.00 (0.00)
	ALASSO-RE	0.934 (0.992)	0.78 (1.00)	0.00 (0.00)
	ALASSO-IC_Q	0.603 (0.237)	0.02 (0.11)	0.00 (0.00)
	True	0.546 (0.218)	0.00 (0.00)	0.00 (0.00)
$n = 200$, $\sigma = 1$	SCAD-RE	0.902 (0.833)	0.55 (1.00)	0.00 (0.00)

β Estimate (D Estimate)				
Model	**Method**	**MRME**	**$ZERO_1$**	**$ZERO_2$**
	SCAD-IC_Q	0.853 (0.487)	0.01 (0.12)	0.00 (0.00)
	LASSO-RE	0.998 (0.998)	0.99 (1.00)	0.00 (0.00)
	LASSO-IC_Q	0.873 (0.554)	0.07 (0.20)	0.00 (0.00)
	ALASSO-RE	0.982 (0.991)	0.79 (1.00)	0.00 (0.00)
	ALASSO-IC_Q	0.871 (0.468)	0.02 (0.11)	0.00 (0.00)
	True	0.839 (0.408)	0.00 (0.00)	0.00 (0.00)

Table 2

Maximum penalized likelihood estimates of Yale infant grown data comparing SCAD and ALASSO penalty functions with random effects and IC_Q penalty estimates

Fixed Estimate[a] (Variance Estimate of Random Effect[b])					
		SCAD		ALASSO	
Variable	**MLE[c]**	**RE**	**IC_Q**	**RE**	**IC_Q**
Intercept	7.002* (-)	6.924 (-)	6.988 (-)	6.913 (-)	6.913 (-)
Visit	2.641* (0.230*)	2.576 (0.087)	2.617 (0.109)	2.543 (0.040)	2.548 (0.067)
Age	−0.035 (0.017)	0.000 (0.000)	0.000 (0.007)	0.000 (0.000)	0.000 (0.000)
Gestation	0.528* (0.017)	0.424 (0.000)	0.455 (0.011)	0.322 (0.000)	0.424 (0.000)
Race	−0.060 (-)	0.000 (-)	0.000 (-)	0.000 (-)	0.000 (-)
Pregnant	−0.004 (-)	0.000 (-)	0.000 (-)	0.000 (-)	0.000 (-)
Gender	0.139* (-)	0.022 (-)	0.033 (-)	0.000 (-)	0.000 (-)
Cocaine	0.103* (-)	0.016 (-)	0.022 (-)	0.000 (-)	0.000 (-)
σ^2 [d]	0.512 (-)	0.552 (-)	0.527 (-)	0.612 (-)	0.594 (-)
IC_Q[e]	9223.7	11507.32	9660.013	11999.01	11773.25

[a] is estimate of **β**

[b] is estimate of diag(**D**)

[c] * indicates significant effects by MLE analysis

[d] is the variance estimate of error term of the linear mixed model

[e] is a measure of goodness of fit